Block clustering of Binary Data with Gaussian Co-variables

Serge Iovleff, Seydou Nourou Sylla, Cheikh Loucoubar

Block clustering of Binary Data with Gaussian Co-variables

Serge Iovleff[1] , Seydou Nourou Sylla[2] and Cheikh Loucoubar[2],
[1] University of Lille, France, serge.iovleff@univ-lille.f
[2] G4-Bio-Informatique - Institut Pasteur, Dakar, Senegal,
seydou.sylla@pasteur.sn, cheikh.loucoubar@pasteur.sn

October 1, 2020

Abstract

The simultaneous grouping of rows and columns is an important technique that is increasingly used in large-scale data analysis. Motivated by the analysis of an epidemiological and genetic data set, we present a novel co-clustering method using co-variables in its construction. It is based on a latent block model taking into account the problem of grouping variables and clustering individuals by integrating information given by a set of co-variables. Numerical experiments on simulated data sets and results on the genetic data set highlight the interest of this approach.

1 Introduction

Clustering is a data analysis method that aims to group together a set of observations into homogeneous classes. It plays an increasingly important role in many scientific and technical fields. Its aim is the automatic problems solving by decision-making based on the observations and to define rules for classifying objects depending on qualitative or quantitative variables. Clustering is the most popular technique for data analysis in many disciplines. In recent years, co-clustering has been increasingly used.

*This work was supported by LIRIMA, International Laboratory for Computer Sciences and Applied Mathematics, Simerge Team.

Unlike classical clustering, which groups similar objects from a single collection of objects, co-clustering or bi-clustering Madeira and Oliveira (2004) aims to group two types of entities simultaneously, based on similarity of their pairwise interactions. It is most often used with bipartite spectral graphing partitioning methods in the field of extracting text data Dhillon (2001) by simultaneously grouping documents and content (words). It is used for analyzing huge corpora of unlabeled documents Xu, Zong, Dolog, and Zhang (2010) in order to simultaneously understand aggregates of subsets of web users sessions and information from the page views. Co-clustering algorithms have also been developed for computer vision applications: it is used for grouping images simultaneously with their low-level visual characteristics and for content-based search Guan, Qiu, and Xue (2005).

In this paper we extend co-clustering methods allowing simultaneous detection of associations between variables and individuals by taking into account covariables. Our method is to be used when one want to co-cluster a set $\mathbf{X}$ of (binary) variables, and individuals in coherence with independent (continuous) variables $\mathbf{Y}$ measured on these same individuals.

This co-clustering approach is motivated by a malaria data set from Senegalese populations (Trape, Tall, Sokhna et al. (2014)). We want to co-cluster a set $\mathbf{X}$ of binary variables, i.e. presence of genetic variants in SNPs (Single-Nucleotide Polymorphism), and individuals (patients) in coherence with a quantitative variable Y measured on these same individuals. In the biological application we have in mind, the additional measure of interest Y should be taken into account by the co-clustering process in order to obtain significant results.

In biology, the selection in a set of variable $\mathbf{X}$ those associated with a quantitative outcome Y, is generally done using genome-wide association studies (GWAS). In a typical GWAS there is measure of hundreds of thousands, or millions, of genetic variants (SNPs), and the attempt is to identify regions harboring SNPs that affect some phenotype of interest, see for example (Timmann, Thye, Vens, Evans, May, Ehmen, Sievertsen, Muntau, Ruge, Loag et al. (2012)) for a GWAS applied to severe falciparum malaria in patients and controls from Ghana, West Africa. This goal can naturally be cast as a variable selection regression problem, with the SNPs as the covariates in the regression. Using the co-clustering presented here, the problem is reversed: we find cluster of individuals and SNPs using the phenotype Y as prior and then find the most influential genetic variants by inspecting the posterior distribution (see parts 2.6 and 3.2).

A second goal in our study is to dichotomize the initially quantitative outcome Y into a binary variable reflecting two categories of risk or two different levels of severity of the disease. Usually, the median value on Y or the value 0 are used as threshold (Loucoubar, Grant, Bureau, Casademont, Bar, Bar-Hen, Diop, Faye, Sarr, Badiane, Tall, Trape, Cliquet, Schwikowski, Lathrop, Paul, and Sakuntabhai

(2016)).

Using, the co-clustering approach presented here, the outcome variable (Y) is automatically clustered in two groups and this allow us to find the optimal cut off to its binarization, i.e. the partition of Y that maximize its association with determinant features.

The paper is organized in two parts. In the first part (section 2) we develop the models, formulas and algorithms. We explain the principle of block mixture models through section 2.1. The latent block model for binary variable taking into account continuous co-variables and the model parameters estimation are proposed in section 2.2. The parameter estimation method is described in section 2.3 and 2.4. The choice of the optimal number of blocks and the measure of influence of each variable on the co-variable $\mathbf{Y}$ are presented in section 2.5 and 2.6. In the second part (section 3) the method is illustrated on simulated data (section 3.1) and on real genetic data (section 3.2).

2 Block mixture models

2.1 Classical latent block model

Let $\mathbf{x}$ be a data set doubly indexed by a set I with n elements (individuals) and a set J with m elements (variables). We represent a partition of I into g clusters by $\mathbf{z} = (z_{11}, \ldots, z_{ng})$ with $z_{ik} = 1$ if i belongs to cluster k and $z_{ik} = 0$ otherwise, $z_i = k$ if $z_{ik} = 1$ and we denote by $z_{.k} = \sum_i z_{ik}$ the cardinality of row cluster k. Similarly, we represent a partition of J into d clusters by $\mathbf{w} = (w_{11}, \ldots, w_{md})$ with $w_{j\ell} = 1$ if j belongs to cluster ℓ and $w_{j\ell} = 0$ otherwise, $w_j = \ell$ if $w_{j\ell} = 1$ and we denote by $w_{.\ell} = \sum_j w_{j\ell}$ the cardinality of column cluster ℓ.

The block mixture model formulation is defined in Govaert and Nadif (2003) and Bhatia, Iovleff, and Govaert (2017) (among others) by the following probability density function

$$f(\mathbf{x};\theta) = \sum_{\mathbf{u}\in\mathscr{U}} p(\mathbf{u};\theta) f(\mathbf{x}|\mathbf{u};\theta)$$

where $\mathscr{U}$ denotes the set of all possible labels of $I \times J$ and θ contains all the unknown parameters of this model. By restricting this model to a set of labels of $I \times J$ defined by a product of labels of I and J, and further assuming that the labels of I and J are independent of each other, one obtains the decomposition

$$f(\mathbf{x};\theta) = \sum_{(\mathbf{z},\mathbf{w})\in\mathscr{Z}\times\mathscr{W}} p(\mathbf{z};\theta) p(\mathbf{w};\theta) f(\mathbf{x}|\mathbf{z},\mathbf{w};\theta) \tag{1}$$

where $\mathscr{Z}$ and $\mathscr{W}$ denote the sets of all possible labellings $\mathbf{z}$ of I and $\mathbf{w}$ of J. Equation (1) defines a *Latent Block Model* (LBM).

2.2 LBM for binary variables with co-variables: General formulation

From now, we assume that $\mathbf{x}$ is a binary data set. Let $\mathbf{y}$ represent a data-set (co-variables) of $\mathbb{R}^p$ indexed by I. In order to take into account this set of co-variables, the classical block model formulation is extended to propose a block mixture model defined by the following probability density function

$$f(\mathbf{x},\mathbf{y};\theta) = \sum_{(\mathbf{z},\mathbf{w})\in\mathscr{Z}\times\mathscr{W}} p(\mathbf{z};\theta)p(\mathbf{w};\theta)f(\mathbf{x},\mathbf{y}|\mathbf{z},\mathbf{w};\theta). \tag{2}$$

By extending the latent class principle of local independence to our block model, data pairs $(x_{ij},\mathbf{y}_i)$, $i=1,\ldots,m$, $j=1,\ldots,n$ are independent once $\mathbf{z}$ and $\mathbf{w}$ are fixed. Hence we have

$$f(\mathbf{x},\mathbf{y}|\mathbf{z},\mathbf{w};\theta) = \prod_{i,j} f(x_{ij},\mathbf{y}_i|z_i,w_j;\theta).$$

We choose to model the dependence between x_{ij} and $\mathbf{y}_i$ using the canonical link for binary response data

$$\begin{aligned} f(x_{ij}|\mathbf{y}_i,\beta_{z_iw_j}) &= \text{logis}(\beta_{0,z_iw_j}+\beta^T_{z_iw_j}\mathbf{y}_i)^{x_{ij}}\left(1-\text{logis}(\beta_{0,z_iw_j}+\beta^T_{z_iw_j}\mathbf{y}_i)\right)^{1-x_{ij}} \\ &= \frac{e^{x_{ij}(\beta_{0,z_iw_j}+\beta^T_{z_iw_j}\mathbf{y}_i)}}{1+e^{\beta_{0,z_iw_j}+\beta^T_{z_iw_j}\mathbf{y}_i}} \end{aligned} \tag{3}$$

with $(\beta_0,\beta_{k,l})\in\mathbb{R}^{p+1}$ and $\text{logis}(x)=e^x/(1+e^x)$. Data points $\mathbf{y}_i$, $i=1,\ldots,m$ are independent once $\mathbf{z}$ is fixed. In the examples presented in section 3, we choose

$$f(\mathbf{y}|\mathbf{z};\theta) = \prod_i \phi(\mathbf{y}_i;\mu_{z_i},\Sigma_{z_i})$$

with ϕ denoting the multivariate Gaussian density in $\mathbb{R}^p$.

In order to simplify the notation, we add a constant coordinate 1 to vectors $\mathbf{y}_i$ and write $\beta_{k,l}$ in the latter rather than $(\beta_{0,k,l},\beta_{k,l})$.

The parameters are thus $\theta=(\pi,\rho,\beta,\mu,\Sigma)$, where $\pi=(\pi_1,\ldots,\pi_g)$ and $\rho=(\rho_1,\ldots,\rho_d)$ are the vectors of probabilities π_k and ρ_ℓ that a row and a column belong to the kth row component and to the ℓth column component respectively, $\beta=(\beta_{kl})$ are the coefficients of the logistic function, μ and Σ are the means and variances of

the Gaussian density. In summary, we obtain the latent block mixture model with pdf

$$f(\mathbf{x},\mathbf{y}|\theta) = \sum_{(\mathbf{z},\mathbf{w})\in\mathscr{Z}\times\mathscr{W}} \prod_{i,j} \pi_{z_i}\rho_{w_j} \frac{e^{x_{ij}(\mathbf{y}_i^T\beta_{z_iw_j})}}{1+e^{\mathbf{y}_i^T\beta_{z_iw_j}}} \phi(\mathbf{y}_i;\mu_{z_i},\Sigma_{z_i}).$$

Using the above expression, the randomized data generation process can be described by the four step row labelling (R), column labelling (C), co-variable data generation (Y) and data generation (X) as follows:

(R) Generate labels $\mathbf{z} = (z_1,\ldots,z_n)$ according to the distribution $\pi = (\pi_1,\ldots,\pi_g)$.
(C) Generate labels $\mathbf{w} = (w_1,\ldots,w_m)$ according to the distribution $\rho = (\rho_1,\ldots,\rho_d)$.
(Y) Generate for $i = 1,...,n$ vector $\mathbf{y}_i$ according to the Gaussian distribution

$$\mathscr{N}_p(\mu_{z_i},\Sigma_{z_i}).$$

(X) Generate for $i = 1,...,n$ and $j = 1,...,m$ a value x_{ij} according to the Bernoulli distribution $f(x_{ij}|\mathbf{y}_i;\beta_{z_iw_j})$ given in (3).

2.3 Model Parameter Estimation

The complete data is represented as a vector $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w})$ where unobservable vectors $\mathbf{z}$ and $\mathbf{w}$ are the labels. The log-likelihood to maximize is

$$l(\theta) = \log f(\mathbf{x},\mathbf{y};\theta) \tag{4}$$

and the double missing data structure, namely $\mathbf{z}$ and $\mathbf{w}$, makes statistical inference more difficult than usual. More precisely, if we try to use an EM algorithm as in standard mixture model Dempster, Laird, and Rubin (1997) the complete data log-likelihood is found to be

$$L_C(\mathbf{z},\mathbf{w},\theta) = \sum_k z_{.k}\log\pi_k + \sum_\ell w_{.\ell}\log\rho_\ell + \sum_{i,j,k,\ell} z_{ik}w_{j\ell}\log f(x_{ij},\mathbf{y}_i;\theta_{k\ell}). \tag{5}$$

The EM algorithm maximizes the log-likelihood $l(\theta)$ iteratively by maximizing the conditional expectation $Q(\theta,\theta^{(c)})$ of the complete data log-likelihood given a previous current estimate $\theta^{(c)}$ and $(\mathbf{x},\mathbf{y})$:

$$\begin{aligned} Q(\theta,\theta^{(c)}) &= \mathbb{E}\left[L_C(\mathbf{z},\mathbf{w},\theta)\,\middle|\,\mathbf{x},\mathbf{y},\theta^{(c)}\right] \\ &= \sum_{i,k} t_{ik}^{(c)}\log\pi_k + \sum_{j,\ell} r_{j\ell}^{(c)}\log\rho_\ell + \sum_{i,j,k,\ell} e_{ikj\ell}^{(c)}\log f(x_{ij},\mathbf{y}_i;\theta_{k\ell}) \end{aligned}$$

where

$$t_{ik}^{(c)} = P(z_{ik} = 1|\mathbf{x}, \mathbf{y}, \theta^{(c)}), \qquad r_{jl}^{(c)} = P(w_{j\ell} = 1|\mathbf{x}, \mathbf{y}, \theta^{(c)}),$$
$$e_{ikj\ell}^{(c)} = P(z_{ik}w_{j\ell} = 1|\mathbf{x}, \mathbf{y}, \theta^{(c)}).$$

Unfortunately, difficulties arise due to the dependence structure in the model, in particular to determine $e_{ikj\ell}^{(c)}$. The assumed independence of $\mathbf{z}$ and $\mathbf{w}$ in (1) is not conserved by the posterior probability.

To solve this problem an approximate solution is proposed in Govaert and Nadif (2003) using the Hathaway (1986) and Neal and Hinton (1998) interpretation of the VEM algorithm. Consider a family of probability distribution $q(z_{ik}, w_{j\ell})$ verifying $q(z_{ik}, w_{j\ell}) > 0$ and the relation $q(z_{ik}, w_{j\ell}) = q(z_{ik})q(w_{j\ell})$, for all i, j, k, l. Set $t_{ik} = q(z_{ik})$ and $r_{jl} = q(w_{j\ell})$, $\mathbf{t} = (t_{ik})_{ik}$ for $i = 1, \ldots, n$, $k = 1, \ldots, g$ and $\mathbf{r} = (r_{jl})_{jl}$ for $j = 1, \ldots, m$ and $l = 1, \ldots, d$. One easily shows that

$$l(\theta) = \tilde{F}_C(\mathbf{t}, \mathbf{r}; \theta) + KL(q(\mathbf{z}, \mathbf{w}) \parallel p(\mathbf{z}, \mathbf{w}|\mathbf{x}, \mathbf{y}, \theta)) \tag{6}$$

with $KL(q \parallel p)$ denoting the Kullback-Liebler divergence of distribution p and q,

$$\tilde{F}_C(\mathbf{t}, \mathbf{r}; \theta) = \sum_k t_{.k} \log \pi_k + \sum_\ell r_{.\ell} \log \rho_l + \sum_{i,j,k,\ell} t_{ik} r_{j\ell} \log f(x_{ij}, \mathbf{y}_i; \theta_{k\ell}) + H(\mathbf{t}) + H(\mathbf{r}) \tag{7}$$

and $H(\mathbf{t})$, $H(\mathbf{r})$ denoting the entropy of $\mathbf{t}$ and $\mathbf{r}$, i.e.

$$H(\mathbf{t}) = \sum_{ik} t_{ik} \log t_{ik}, \qquad H(\mathbf{r}) = \sum_{jl} r_{jl} \log r_{jl}.$$

$\tilde{F}_C$ is called the free energy or the fuzzy criterion. As the Kullback-Liebler divergence is always positive, the fuzzy criterion is a lower bound of the log-likelihood and is used as a replacement for it. Doing that, the maximization of the likelihood $l(\theta)$ is replaced by the following problem

$$\underset{\mathbf{t}, \mathbf{r}, \theta}{\operatorname{argmax}} \tilde{F}_C(\mathbf{t}, \mathbf{r}, \theta).$$

This maximization can be achieved using the BEM algorithm detailed hereafter.

2.4 Block expectation maximization (BEM) Algorithm

The fuzzy clustering criterion given in (7) can be maximized using a variational EM algorithm (VEM). We here outline the various expressions evaluated during E and M steps.

E-Step: we compute either the values of $\mathbf{t}$ (respectively $\mathbf{r}$) with $\mathbf{r}$ (respectively $\mathbf{t}$) and θ fixed (formulas (11), (12) hereafter). Details are given in appendix A.

M-Step: we calculate row proportions π and column proportions ρ. The maximization of $\tilde{F}_C$ w.r.t. π, and w.r.t ρ, is obtained by maximizing $\sum_k t_{.k} \log \pi_k$, and $\sum_\ell r_{.\ell} \log \rho_\ell$ respectively, which leads to

$$\pi_k = \frac{t_{.k}}{n} \quad \text{and} \quad \rho_\ell = \frac{r_{.\ell}}{m}. \tag{8}$$

Also, for $\mathbf{t}$, $\mathbf{r}$ fixed, the estimate of model parameters β is obtained by maximizing

$$\beta_{kl} = \underset{\beta}{\operatorname{argmax}} \sum_{ij} t_{ik} r_{jl} \log f(x_{ij}|\mathbf{y}_i;\beta), \quad k = 1,\ldots,g, \quad l = 1,\ldots,d. \tag{9}$$

Details are given in appendix B. Finally parameters of the Gaussian density are given by the usual formulas

$$\mu_k = \frac{1}{t_{.k}} \sum_i t_{ik} \mathbf{y}_i \qquad \text{and} \qquad \Sigma_k = \frac{1}{t_{.k}} \sum_i t_{ik} (\mathbf{y}_i - \mu_k)(\mathbf{y}_i - \mu_k)^T. \tag{10}$$

Putting everything together, we obtain the **BEM** algorithm.

BEM algorithm: Using the **E** and **M** steps defined above, **BEM** algorithm can be decomposed as follows:

Initialization Set $\mathbf{t}^{(0)}, \mathbf{r}^{(0)}$ and $\theta^{(0)} = (\pi^{(0)}, \rho^{(0)}, \beta^{(0)}, \mu^{(0)}, \Sigma^{(0)})$.
(a) Row-EStep Compute $\mathbf{t}^{(c+1)}$ using formula

$$t_{ik}^{(c+1)} = \frac{\pi_k^{(c)} \prod_{jl} \left(f(x_{ij}|\mathbf{y}_i;\beta_{kl}^{(c)}) \phi(\mathbf{y}_i;\mu_k^{(c)},\Sigma_k^{(c)}) \right)^{r_{jl}^{(c)}}}{\sum_k \pi_k^{(c)} \prod_{jl} \left(f(x_{ij}|\mathbf{y}_i;\beta_{kl}^{(c)}) \phi(\mathbf{y}_i;\mu_k^{(c)},\Sigma_k^{(c)}) \right)^{r_{jl}^{(c)}}}. \tag{11}$$

(b) Row-MStep Compute $\pi^{(c+1)}$, $\mu^{(c+1)}$, $\Sigma^{(c+1)}$ using equations (8) and (10) and estimate $\beta^{(c+1/2)}$ by solving maximization problem (9).
(c) Col-EStep Compute $\mathbf{r}^{(c+1)}$ using formula

$$r_{jl}^{(c+1)} = \frac{\rho_l^{(c)} \prod_{ik} f(x_{ij}|\mathbf{y}_i;\beta_{kl}^{(c+1/2)})^{t_{ik}^{(c+1)}}}{\sum_l \rho_l^{(c)} \prod_{ik} f(x_{ij}|\mathbf{y}_i;\beta_{kl}^{(c+1/2)})^{t_{ik}^{(c+1)}}}. \tag{12}$$

Observe that r_{jl} does not depend of the density of $\mathbf{y}$.

(d) Col-MStep Compute $\rho^{(c+1)}$ using equations (8) and estimate $\beta^{(c+1)}$ by solving maximization problem (9).

Iterate (a)-(b)-(c)-(d) until convergence.

2.5 Selecting the number of blocks

BIC is an information criterion defined as an asymptotic approximation of the logarithm of the integrated likelihood (Schwarz et al. (1978)). The standard case leads to write BIC as a penalised maximum likelihood:

$$\text{BIC} = -2\max_{\theta} l(\theta) + D\log(N)$$

where N is the number of statistical units and D the number of free parameters and $l(\theta)$ defined in (4). Unfortunately, this approximation cannot be used for LBM, due to the dependency structure of the observations $(\mathbf{x},\mathbf{y})$. However, a heuristic have been stated to define BIC in Keribin, Brault, Celeux, and Govaert (2012) and Keribin, Brault, Celeux, and Govaert (2015). BIC-like approximations ICL lead to the following approximation as n and m tend to infinity

$$\text{BIC}(g,d) = -2\max_{\theta} \log f(\mathbf{x},\mathbf{y};\theta) + (g-1)\log n + (d-1)\log m + \lambda \log n + gd(p+1)\log(mn)$$

with λ the number of parameters of the $\mathbf{y}$ distribution. For LBM, the intractable likelihood $f(\mathbf{x},\mathbf{y};\theta)$ is replaced by the maximized free energy $\tilde{F}_C$ in (7) obtained by the BEM algorithm.

2.6 Measuring the Influence of a Variable

Let j be fixed (a column of the matrix $\mathbf{x}$). We would like to measure the effect of the variable $\mathbf{x}^j = (x_{ij})_{i=1}^n$ on $\mathbf{y}$. It is possible to obtain a measure of this effect by looking to the posterior probability of $\mathbf{y}$.

Lemma 1 *Let $(\mathbf{x},\mathbf{z},\mathbf{w})$ be fixed. For $l = 1,\ldots,d$ let m_l denotes the number of columns with label l, i.e $m_l = \#\{w_{jl} = 1,\ j = 1,\ldots m\}$ and for a row i fixed let m_{il}*

denotes the number of elements such that $w_{jl} = 1$ *and* $x_{ij} = 1$*, i.e.* $m_{il} = \#\{w_{jl}x_{ij} = 1,\ j = 1,\dots n\}$*. The posterior probability of the co-variable* $\mathbf{y}$ *is*

$$\begin{aligned} f(\mathbf{y}|\mathbf{x},\mathbf{z},\mathbf{w},\theta) &\propto \prod_{i=1}^{n}\prod_{l=1}^{d} \pi_{z_i}\rho_l^{m_l}\,\mathrm{logis}(\mathbf{y}_i^T\beta_{z_i l})^{m_{il}}\left(1-\mathrm{logis}(\mathbf{y}_i^T\beta_{z_i l})\right)^{m_l-m_{il}} \\ &\quad \phi(\mathbf{y}_i;\mu_{z_i},\Sigma_{z_i}) \\ &\propto \prod_{i=1}^{n}\pi_{z_i}\phi(\mathbf{y}_i;\mu_{z_i},\Sigma_{z_i})\prod_{l=1}^{d}\rho_l^{m_l}\frac{e^{m_{il}\mathbf{y}_i^T\beta_{z_i l}}}{\left(1+e^{\mathbf{y}_i^T\beta_{z_i l}}\right)^{m_l}} \end{aligned}$$

Alternatively, for $k = 1,\dots,g$*, let* n_k *denotes the number of rows with label k, i.e.* $n_k = \#\{z_{ik} = 1, i = 1,\dots,n\}$*. The posterior probability of the co-variable* $\mathbf{y}$ *is*

$$f(\mathbf{y}|\mathbf{x},\mathbf{z},\mathbf{w},\theta) \propto \prod_{j=1}^{m}\rho_{w_j}\prod_{k=1}^{g}\pi_k^{n_k}\prod_{i:z_i=k}\frac{e^{x_{ij}(\mathbf{y}_i^T\beta_{z_i w_j})}}{1+e^{\mathbf{y}_i^T\beta_{z_i w_j}}}\phi(\mathbf{y}_i;\mu_k,\Sigma_k).$$

The proof of this lemma is straightforward and therefore omitted.

Assuming $\mathbf{z}$ and $\mathbf{w}$ known, we measure the influence of a variable using its contribution to the posterior probability. Fixing j, taking the logarithm and eliminating terms independent of $\mathbf{x}^j$, we obtain the *influence measure criteria*

$$I(j) = \log\rho_{w_j} + \sum_{i=1}^{n}\left(x_{ij}\mathbf{y}_i^T\beta_{z_i w_j} - \log(1+\exp(\mathbf{y}_i^T.\beta_{z_i w_j}))\right). \tag{13}$$

Equation (13) is interpreted as the log-contribution to the posterior distribution of the variable $\mathbf{x}^j$. Replacing the unknown labels w_j and z_i by their MAP estimators $\hat{w}_j$ and $\hat{z}_i$, we are able to sort the variables from the most to the less influential.

3 Numerical Experiments

3.1 Simulated data

In this part, we fix $g = 2$ (the number of cluster of the individuals) and we study the number of correctly classified rows/columns for different values of m, n, d. This study is only partial and is focused to values "near" the dimensions of the real data set which motivate this work (section 3.2).

We also conduct experiments about the computational time when the number of columns is huge. These experiments were performed for a future work and are only given for information. The real data set illustrating this paper has a much less number of columns.

3.1.1 Error rate

We simulate 80 times data set and average the number of columns correctly classified The cluster of a column is estimated using the maximum a posterior (MAP) estimator

$$\hat{w}_j = \arg\max_{l=1}^{d} r_{jl}.$$

Comparison between the graphics in figure 1 shows that the number of incorrectly classified columns labels increases as d increases while it remains relatively constant with m. An other salient feature is that when the number of individuals m is greater, this error rate is lower. The number of correctly classified rows is stable near 0.9 for all tested configurations of the parameters and is not displayed.

3.1.2 Computational time

We compute 80 times the elapsed time of the estimation procedure for various configurations of the parameters on a HP Zbook G3 with Intel Core i7-6700HQ (2.60 GHz, 2133 MHz, 6 MB L3 cache, 4 cores, 45W). The (averaged) computing time as a function of m when $g = 2$ for different values of m (the number of columns) and when d (the number of cluster in columns) take values 2 and 6 is plotted in figure 2 below We can observe that as n grows the elapsed time grows linearly, but that the slope increases as d (the number of class in columns) is increased.

3.2 Real Data Analysis

Here, we study data from an epidemiological and genetic survey of malaria disease in Senegal. Data were collected between 1990 to 2008. We worked on a dataset including $n = 885$ individuals with measured malaria risk score (phenotype) and with genotypes available on several candidate genes for susceptibility/resistance to the disease. A total of $m = 45$ Single Nucleotide Polymorphisms (SNPs) were considered across these genes and were used as genetic variables. The malaria risk score was a quantitative measure normally distributed and was considered as a covariable for this co-clustering method. The SNPs are coded in dominant effect on the disease risk. Using the BIC criteria (figure 3), we choose to focus on the model with $d = 2$ groups of individuals and $g = 11$ groups of SNPs.

3.2.1 Analysis for phenotype data

We used the malaria risk score as phenotype. The choice of a mixture model or not depends on the application context.

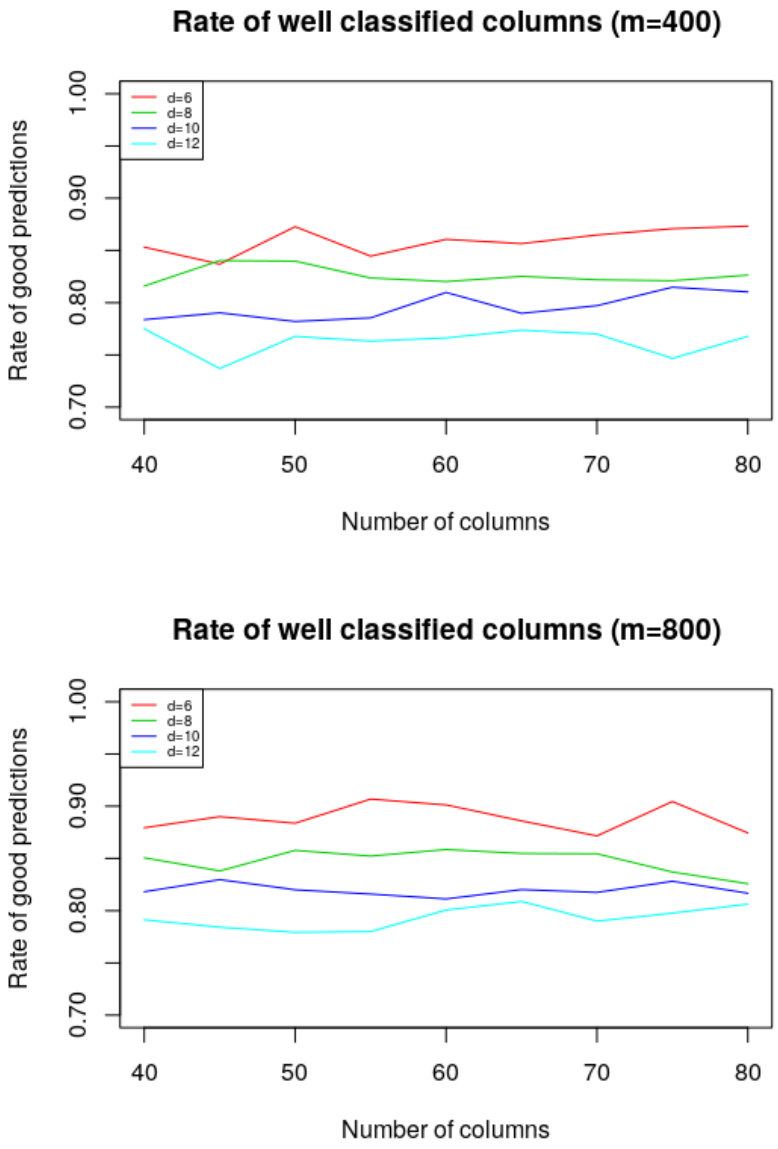


Figure 1: Rates of correctly classified columns when the number of rows is 400 and 800. The number of columns (m) is between 40 and 80. The number of cluster (d) is between 6 and 12. There is only two groups of rows.

In the case of genetic epidemiology, we are often interested in the comparison between the *susceptible* and the *resistant* to a given disease. Here, we looked

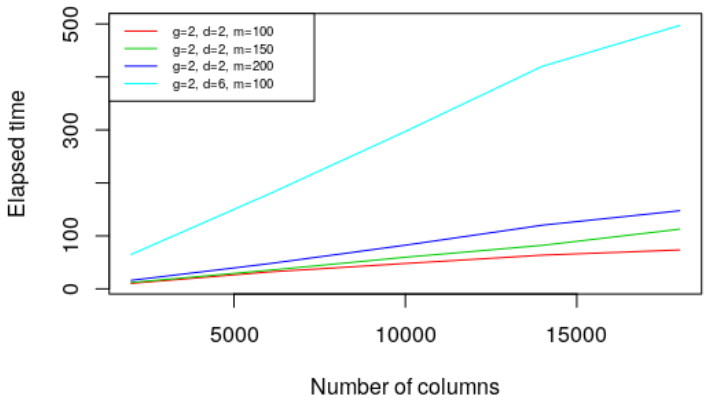


Figure 2: computational elapsed time for $m = 2000$, 6000, 10000, 14000 and 18000 (in minutes) and for various values of n.

for genes that could explain differences between *susceptible* and *resistant*, justifying the use of a mixture model on this phenotype.

After block-clustering, we found that individuals were divided into two groups: Individuals with lowest malaria risk scores (-2.101 to 0.040) defined as *resistant* and individuals with highest malaria risk scores values (0.042 to 3.509) defined as *susceptible*.

In figure 4 we observe on the left a bi-modal distribution of the phenotype after clustering (i.e. when $\mathbf{y}$ is conditioned by $(\mathbf{x}, \mathbf{z})$). On the right we observe how SNP with high/low levels of mutations are grouped together.

3.2.2 Analysis for genotypes data

We looked at the SNPs to determine which ones would potentially be involved in malaria susceptibility / resistance.

The proposed methodology allowed the selection of the most significant SNPs according to the influence measure proposed in section 2.6. The most influential SNPs belong to class 1 and class 9. It is noted that the SNPs of these classes have been shown in the literature to have a high significance effect on malaria. Most Glucose-6-Phosphate Dehydrogenase (G6PD) and hemoglobin SNPs are grouped into these 2 classes. Reviews from exiting literature give us: G6PD deficiency is

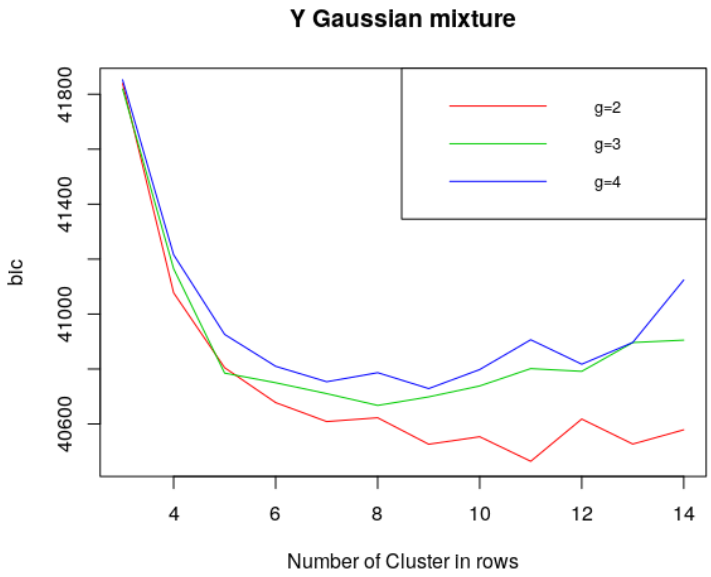
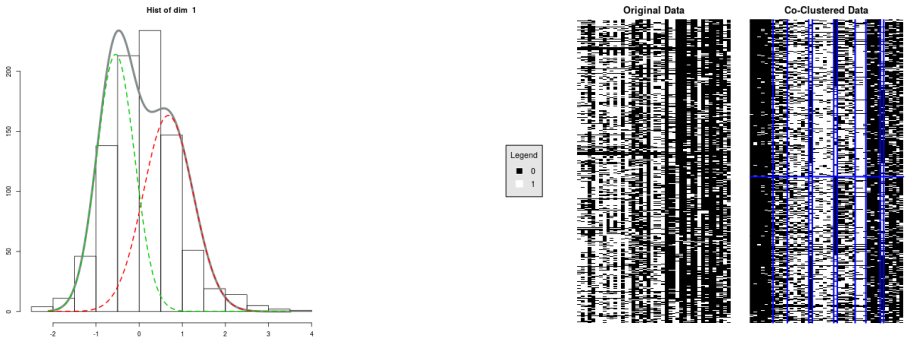


Figure 3: BIC computation for different values of d and g. We observe that it is minimal for $g = 2$ and $d = 11$ among tested d values $(2, \ldots, 14)$ and g values $(2, 3, 4)$.

prevalent in sub-Saharan African populations and has been associated with protection against severe malaria Maiga, Dolo, Campino, Sepulveda, Corran, Rockett, Troye-Blomberg, Doumbo, and Clark (2014), Manjurano, Sepulveda, Nadjm, Mtove, Wangai, Maxwell, Olomi, Reyburn, Riley, Drakeley et al. (2015), Toure, Konate, Sissoko, Niangaly, Barry, Sall, Diarra, Poudiougou, Sepulveda, Campino, Rockett, Clark, Thera, Doumbo, and in collaboration with The MalariaGEN Consortium (2012). Hemoglobins S and C (HbS and HbC respectively) are well known to be two variant forms of normal adult hemoglobin (HbA) resulting from distinct mutations in the β-globin gene. The protective effect of HbS against Plasmodium falciparum malaria has been shown by several authors Beet et al. (1946), Allison (1954b,a), Williams (2011). In the case of HbC, the protection is highest in homozygous individuals with HbC.

The results found with our co-clustering method confirm the link between malaria and sickle cell polymorphism (Hbs) and G6PD.



(a) (b)

Figure 4: Co-Clustering Results: phenotype and SNPs.
(a) - Empirical Distribution of the phenotype (histogram) - Distribution of the *susceptible* (red) - distribution of the *resistant* (green) - mixing distribution (grey).
(b) Array with the presence/absence of mutations before and after block-clustering

3.2.3 Association between phenotype and genotypes

The most common approach used to screen association between genetic data and phenotypic data is the GWAS method (Genome Wide Association Studies). GWAS usually performs linear regressions of a quantitative phenotype on each of the genotype variables.

Here, we proposed a co-clustering method, that can state as a step prior to association analysis. The method reorganizes data by simultaneously identifying optimal partitions of the phenotype and clusters of markers so that this partition of the phenotype would be better explained by a given set of markers.

In the co-clustering method, the phenotype is used as co-variable to find optimal clusters on the variables as well as on the individuals. In our Senegalese malaria data, we obtained a dichotomy of the phenotype. This dichotomy allowed us to divide individuals into two categories: *susceptible* and *resistant*. In this part, we compare results of GWAS studies between classical approaches and the one using co-clustering as a first step.

- Classical approaches:
  - testing association between the raw phenotype and genotypes by simple linear regression
  - testing association between the binarized phenotype (by choosing a cut-off, e.g. value zero) and genotypes by logistic regression.
- Using co-clustering followed by association tests:

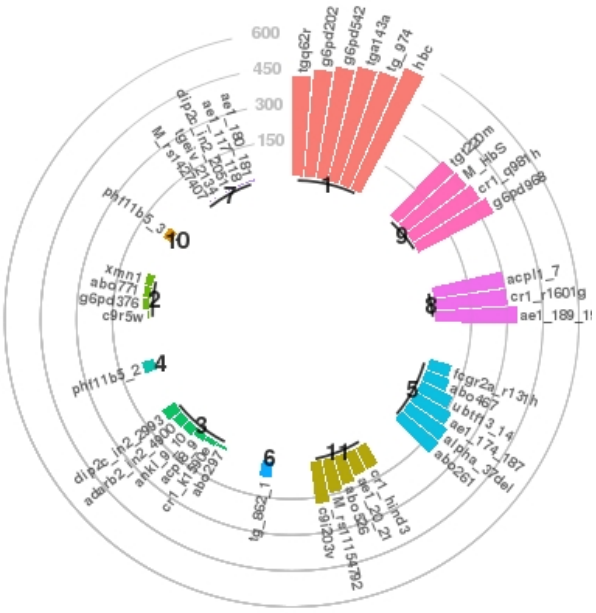


Figure 5: Representation of each block variable according to the influence measure. Here, block 1 has the highest influence measures (417 to 553) while block 7 has the lowest (1 to 6)

- testing association between the binarized phenotype (using the optimal cutoff identified by the co-clustering algorithm, here value 0.042) and genotypes by logistic regression.

The association signal between the phenotype and each of the genotype variables (here the SNPs) is represented by the corresponding p-value. Lower the p-value, better the association signal. Number of good signals (i.e. SNPs with p-values less than 0.05) is compared between each approach (liner regression, logistic regression using 0 as cut off to binarize Y, logistic regression using the optimal cut off identified through co-clustering of the data). It is presented in Figure 6 at different level of significance of the signal (p-value 5e-4 ; 5e-4 ¡ p-value 5e-3 ; 5e-3 ¡ p-value 5e-2).

Figure 6 shows that there is more signals at the 5% threshold for the phenotype from co-clustering compared to the two other phenotypes.

In summary the proposed methodology allows to detect more significant SNPs compared to classical methods by increasing detection power through optimal

clustering of the data.

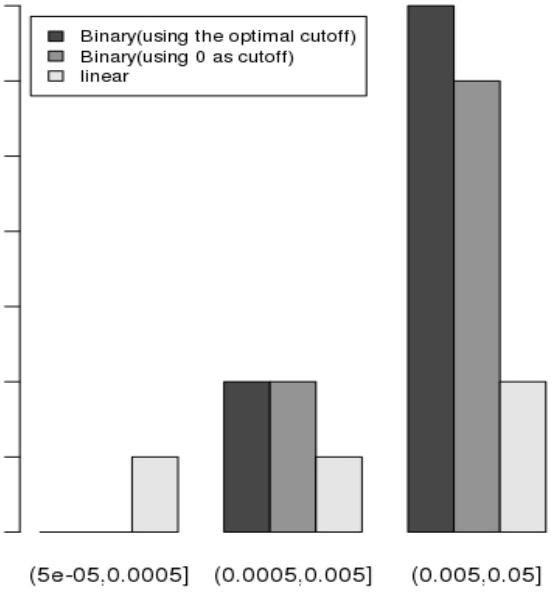


Figure 6: Number of significant p-values of each method

4 Conclusion

In this article, our main contribution is to develop a co-clustering model taking into account a (mixture of) Gaussian co-variables. Applications have been made on simulated and real data sets. Our preliminary results are confirmed in previous studies in Africa. The method offers good classification performance on complex data sets (large number of variables and classes). This method can be useful in a wide variety of classification problems with Gaussian predictors and will allow us to discover new patterns of genes allowing to understand and evaluate the mechanism existing between genetics and malaria in an African population particularly in a Senegalese rural area. Further analysis could be done with more SNPs in another paper in preparation. Estimation is performed using a R package (with computational part in C++) that will be soon be available on the CRAN website `https://cran.r-project.org/`. Meanwhile the package is available on demand to the authors.

References

Allison, A. C. (1954a): "The distribution of the sickle-cell trait in east africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 48, 312–318.

Allison, A. C. (1954b): "Protection afforded by sickle-cell trait against subtertian malarial infection," *British medical journal*, 1(4857), 290–4.

Beet, E. et al. (1946): "Sickle cell disease in the balovale district of northern rhodesia," *East African medical journal*, 23, 75–86.

Bhatia, P., S. Iovleff, and G. Govaert (2017): "blockcluster: An r package for model-based co-clustering," *Journal of Statistical Software, Articles*, 76, 1–24, URL `https://www.jstatsoft.org/v076/i09`.

Dempster, A., N. Laird, and D. Rubin (1997): "Maximum likelihood from incomplete data with the em algorithm (with discussion)," *Journal of the Royal Statistical Society, Series B*, 39, 1.

Dhillon, I. S. (2001): "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, New York, NY, USA: ACM, 269–274, URL `http://doi.acm.org/10.1145/502512.502550`.

Govaert, G. and M. Nadif (2003): "Clustering with block mixture models," *Pattern Recognition*, 36, 463–473, URL `http://www.sciencedirect.com/science/article/pii/S0031320302000742`.

Guan, J., G. Qiu, and X. Xue (2005): "Spectral images and features co-clustering with application to content-based image retrieval," in *2005 IEEE 7th Workshop on Multimedia Signal Processing*, 1–4.

Hathaway, R. J. (1986): "Another interpretation of the em algorithm for mixture distributions," *Statistics & Probability Letters*, 4, 5356, URL `http://www.sciencedirect.com/science/article/pii/0167715286900167`.

Keribin, C., V. Brault, G. Celeux, and G. Govaert (2012): "Model selection for the binary latent block model," in *20th International Conference on Computational Statistics (COMPSTAT 2012)*, Limassol, Cyprus, 379–390.

Keribin, C., V. Brault, G. Celeux, and G. Govaert (2015): "Estimation and selection for the latent block model on categorical data," *Statistics and Computing*, 25, 1201–1216, URL `https://doi.org/10.1007/s11222-014-9472-2`.

Loucoubar, C., A. Grant, J. Bureau, I. Casademont, N. Bar, A. Bar-Hen, M. Diop, J. Faye, F. Sarr, A. Badiane, A. Tall, J.-F. Trape, F. Cliquet, B. Schwikowski, M. Lathrop, R. Paul, and A. Sakuntabhai (2016): "Detecting multi-way epistasis in family-based association studies," *Briefings in Bioinformatics*.

Madeira, S. C. and A. L. Oliveira (2004): "Biclustering algorithms for biological

data analysis: a survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1, 24–45.

Maiga, B., A. Dolo, S. Campino, N. Sepulveda, P. Corran, K. A. Rockett, M. Troye-Blomberg, O. K. Doumbo, and T. G. Clark (2014): "Glucose-6-phosphate dehydrogenase polymorphisms and susceptibility to mild malaria in dogon and fulani, mali," *Malaria Journal*, 13, 270, URL https://doi.org/10.1186/1475-2875-13-270.

Manjurano, A., N. Sepulveda, B. Nadjm, G. Mtove, H. Wangai, C. Maxwell, R. Olomi, H. Reyburn, E. M. Riley, C. J. Drakeley, et al. (2015): "African glucose-6-phosphate dehydrogenase alleles associated with protection from severe malaria in heterozygous females in tanzania," *PLoS genetics*, 11, e1004960.

Neal, R. M. and G. E. Hinton (1998): *A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants*, Dordrecht: Springer Netherlands, 355–368, URL https://doi.org/10.1007/978-94-011-5014-9_12.

Schwarz, G. et al. (1978): "Estimating the dimension of a model," *The annals of statistics*, 6, 461–464.

Timmann, C., T. Thye, M. Vens, J. Evans, J. May, C. Ehmen, J. Sievertsen, B. Muntau, G. Ruge, W. Loag, et al. (2012): "Genome-wide association study indicates two novel resistance loci for severe malaria," *Nature*, 489, 443–446.

Toure, O., S. Konate, S. Sissoko, A. Niangaly, A. Barry, A. H. Sall, E. Diarra, B. Poudiougou, N. Sepulveda, S. Campino, K. A. Rockett, T. G. Clark, M. A. Thera, O. Doumbo, and in collaboration with The MalariaGEN Consortium (2012): "Candidate polymorphisms and severe malaria in a malian population," *PLOS ONE*, 7, 1–6, URL https://doi.org/10.1371/journal.pone.0043987.

Trape, J.-F., A. Tall, C. Sokhna, et al. (2014): "The rise and fall of malaria in a west african rural community, dielmo, senegal, from 1990 to 2012: a 22 year longitudinal study," *Lancet Infect. Dis.*, 14, 476488.

Williams, T. N. (2011): "How do hemoglobins s and c result in malaria protection?" *The Journal of Infectious Diseases*, 204, 1651–1653, URL http://dx.doi.org/10.1093/infdis/jir640.

Xu, G., Y. Zong, P. Dolog, and Y. Zhang (2010): "Co-clustering analysis of weblogs using bipartite spectral projection approach," *Knowledge-Based and Intelligent Information and Engineering Systems*, 398–407.

A Computing the (rows and columns) E-Step

For the E-Step t_{ik} value maximize the fuzzy criterion given in equation (7). Derivative with respect to t_{ik} gives

$$\frac{\partial \tilde{F}_C(\mathbf{t},\mathbf{r};\theta)}{\partial t_{ik}} = \log \pi_k + \sum_{j,\ell} r_{j\ell} \log f_{k\ell}(x_{ij},\mathbf{y}_i;\theta) - \log t_{ik} - 1.$$

Equating this equation to zero, taking exponential and recalling that $\sum_k t_{ik} = 1$, we obtain that t_{ik} is updated as

$$t_{ik}^{(c+1)} = \frac{\pi_k^{(c)} \prod_{j,l} \left[f(x_{ij},\mathbf{y}_i;\theta^{(c)}) \right]^{r_{jl}^{(c)}}}{\sum_k \prod_{j,l} \left[f(x_{ij},\mathbf{y}_i;\theta^{(c)}) \right]^{r_{jl}^{(c)}}}.$$

For numerical reason, we prefer to compute the logarithm of this expression which is

$$\log(t_{ik}^{(c+1)}) \propto \log(\pi_k^{(c)}) + \sum_{j,l} r_{jl}^{(c)} \log f(x_{ij},\mathbf{y}_i;\theta^{(c)}).$$

Recall that (see equation 3)

$$\begin{aligned}
\log f(x_{ij}|\mathbf{y}_i;\beta_{kl}^{(c)}) &= x_{ij} \log(\text{logis}(\mathbf{y}_i^T \beta_{kl}^{(c)})) + (1-x_{ij}) \log(1-\text{logis}(\mathbf{y}_i^T \beta_{kl}^{(c)})) \\
&= \log(1-\text{logis}(\mathbf{y}_i^T \beta_{kl}^{(c)})) + x_{ij} \log\left(\frac{\text{logis}(\mathbf{y}_i^T \beta_{kl}^{(c)})}{1-\text{logis}(\mathbf{y}_i^T . \beta_{kl})} \right) \\
&= \log(1+\exp(\mathbf{y}_i^T \beta_{kl}^{(c)})) + x_{ij} \mathbf{y}_i^T \beta_{kl}^{(c)}
\end{aligned}$$

giving

$$\log t_{ik}^{(c+1)} \propto \log \pi_k^{(c)} + \sum_{j,l} r_{jl}^{(c)} x_{ij} \mathbf{y}_i^T . \beta_{kl}^{(c)} - \sum_l r_{.l}^{(c)} \log(1+e^{\mathbf{y}_i^T . \beta_{kl}^{(c)}}) + m \log \phi(\mathbf{y}_i;\mu_k^{(c)},\Sigma_k^{(c)}).$$

Similar computation gives for r_{jl}

$$\log(r_{jl}^{(c+1)}) \propto \log\left(\rho_l^{(c)}\right) + \sum_{i,k} t_{ik}^{(c+1)} \left(x_{ij} \mathbf{y}_i^T \beta_{kl}^{(c+1/2)} - \log\left(1+e^{\mathbf{y}_i^T . \beta_{kl}^{(c+1/2)}}\right)\right).$$

Observe that the Gaussian distribution does not depend of j nor l. This term become constant when summing over i and k and disappears when r_{jl} values are normalized.

B Computing the M-Step

For the M-Step, we use a Newton-Raphson algorithm in order to solve the equation (9). For each pair (k,l) the function to maximize can be written

$$\ell_{k,l}(\beta) = \sum_{i,j} \left(r_{jl} t_{ik} x_{ij} \mathbf{y}_i^T \beta - r_{jl} t_{ik} \log(1 + \exp(\mathbf{y}_i^T . \beta)) \right)$$

The first derivative with respect to the d-th coordinate β_d is

$$\frac{\partial \ell_{k,l}(\beta)}{\partial \beta_d} = \sum_{i,j} \left(r_{jl} t_{ik} x_{ij} y_{i,d} - r_{jl} t_{ik} y_{i,d} \frac{\exp(\mathbf{y}_i^T \beta)}{1 + \exp(\mathbf{y}_i^T \beta)} \right)$$

giving the following expression for the gradient

$$\nabla_\beta \ell_{k,l}(\beta) = Y^T D (X - \mu)$$

with $Y = [\mathbf{y}_i]_{i=1}^N$, $X = \left[\sum_j r_{jl} x_{ij}\right]_{i=1}^N$, $\mu = \left[r_{.l} \frac{\exp(\mathbf{y}_i^T . \beta)}{1+\exp(\mathbf{y}_i^T \beta)} \right]_{i=1}^N$, $D = \text{diag}(t_{ik})_{i=1}^N$ The second derivative with respect to β_d and $\beta_{d'}$ is

$$\frac{\partial^2 \ell_{k,l}(\beta)}{\partial \beta_d \partial \beta_{d'}} = -\sum_{i,j} \left(r_{jl} t_{ik} y_{i,d} y_{i,d'} \frac{\exp(\mathbf{y}_i^T \beta)}{(1 + \exp(\mathbf{y}_i^T \beta))^2} \right)$$

giving the following expression for the hessian

$$H_\beta = -Y^t D W Y \qquad \text{with} \qquad W = \text{diag}\left(\frac{r_{.l} \exp(\mathbf{y}_i^T . \beta)}{(1 + \exp(\mathbf{y}_i^T \beta))^2} \right) = \text{diag}\left(r_{.l} \, \mu_i (1 - \mu_i) \right)$$